Homework 6

Hasmik Aleksanyan

25 05 2020

Problem 1

(a) What are the differences among exclusive, overlapping and fuzzy clustering? Bring(create your own) an example of fuzzy clustering with k = 2. Use the function funny() from library cluster and data visualization techniques from package factoextra to show your results. Show the membership matrix. Which of your observations belongs to both clusters?

(b) Suppose we have an example of a data set with 20 observations. We need to cluster the data using the K-means algorithm. After clustering using k =1,2,3,4 and 5 we obtained only one non-empty cluster. How is it possible?

(c) Suppose we have an example of a data set consisting of three natural circular clusters. These clusters have the same number of points and have the same distribution. The centers of clusters lie on a line, the clusters are located such that the center of the middle cluster is equally distant from the other two. Why will not bisecting K-means find the correct cluster?

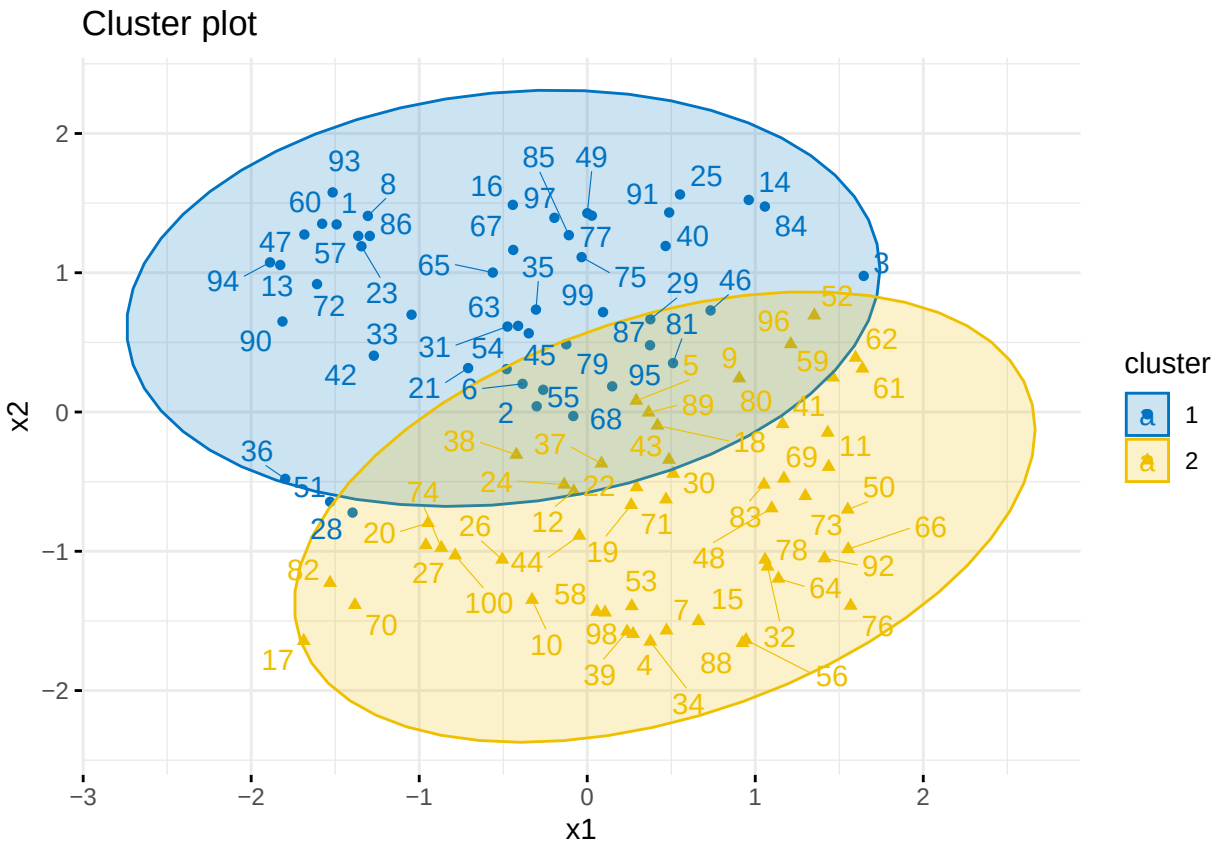
Solution 1

(a) Exclusive clustering - each data object can only exist in one cluster. Overlapping - data objects can exist in any of the clusters. Cause there is no such definite boundary to separate clusters. Fuzzy clustering - every data object belongs to every cluster and its relashionship is defined by the membership weight that is between 0 to 1.

```
set.seed(165)
x1 <- runif(100,0,2)
x2 <-runif(100, 1,3)
dt <- data.frame(x1,x2)


res.fanny <- fanny(dt, 2)
head(res.fanny$membership)
```

```
##           [,1]      [,2]
## [1,] 0.7196431 0.2803569
## [2,] 0.5807396 0.4192604
## [3,] 0.5047675 0.4952325
## [4,] 0.2722097 0.7277903
## [5,] 0.4689677 0.5310323
## [6,] 0.6539022 0.3460978
```

```
fviz_cluster(res.fanny, ellipse.type = "norm", repel = TRUE,
             palette = "jco", ggtheme = theme_minimal(),
             legend = "right")
```



(b) I think that is because all data points belong to anly one cluster.

(c) I think it will split in two clusters, and will take centers between middle cluster center and 2 others.

Problem 2

Consider the following dataset: Table 1: Dataset to Perform K-Means Variable.1 Variable.2 1 2 5 2 2 3 3 8 3 4 0 4 5 7 5 6 0 7 7 1 5 8 7 3 9 3 7 10 9 5

The goal of this task is to perform K-means clustering via R (manually ¨ ˆ), with k =2, using data with 2 features from the table above. Follow the step above:

(a) Neatly plot the observations using ggplot.

(b) Randomly assign a cluster label to each observation. You can use the sample() command in R to do it. Report the cluster labels for each observation.

(c) Define the coordinates of the centroids for each cluster. Show your results.

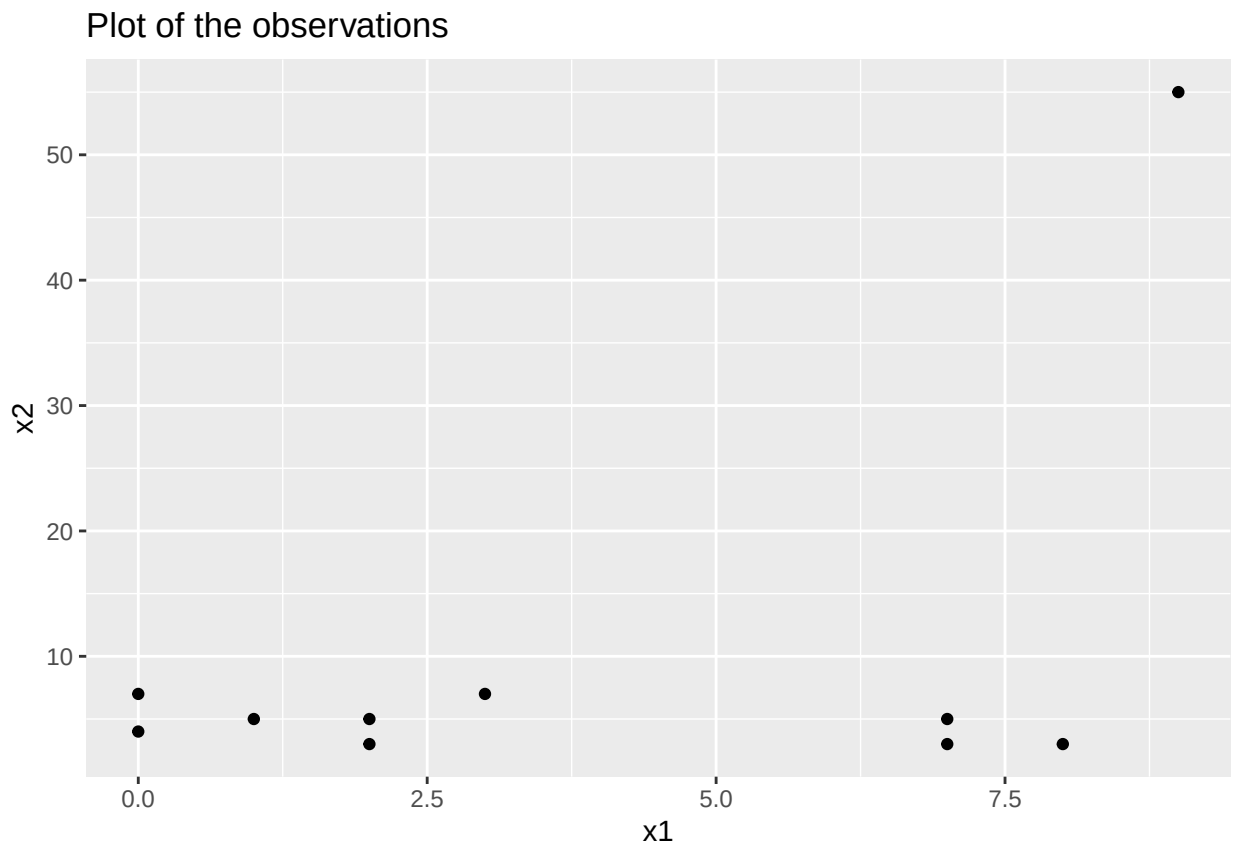
(d) Assign each observation to the cluster using the closeness of each observation to the centroids, in terms of Euclidean distance. Report the cluster labels for each observation.

(e) Repeat (c) and (d) until the centroids remain the same. You can use loops for this task.

(f) Show the observations on the plot by coloring them according to the clusters obtained. Show centroids on the plot.

Solution 2

```
set.seed(151)
x1 <- c(2,2,8,0,7,0,1,7,3,9)
x2 <- c(5,3,3,4,5,7,5,3,7,55)
dat <- data.frame(x1,x2)


dat %>%
  ggplot(aes(x1, x2)) +
  geom_point() +
  ggtitle("Plot of the observations")
```




```
clust <- sample(c(1,2), size = nrow(dat), replace = TRUE)
clust
```

```
##  [1] 2 1 1 2 2 2 2 2 1 1
```

```
(centroid1 <- c(mean(dat[clust == 1, 1]), mean(dat[clust == 1, 2])))
```

```
## [1]  5.5 17.0
```

```r
(centroid2 <- c(mean(dat[clust == 2, 1]), mean(dat[clust == 2, 2])))
```

```
## [1] 2.833333 4.833333
```

```r
distance <- function (x, y){
  return(sqrt((x[1] - y[1])^2 + (x[2] - y[2])^2))
}

for (i in 1:10){
  dis1 <- distance(dat[i,], centroid1)
  dis2 <- distance(dat[i,], centroid2)
  if (dis1 < dis2) {clust[i] <- 1}
  else{clust[i] <- 2}
}

(centroid1 <- c(mean(dat[clust == 1, 1]), mean(dat[clust == 1, 2])))
```

```
## [1]  9 55
```

```r
(centroid2 <- c(mean(dat[clust == 2, 1]), mean(dat[clust == 2, 2])))
```

```
## [1] 3.333333 4.666667
```

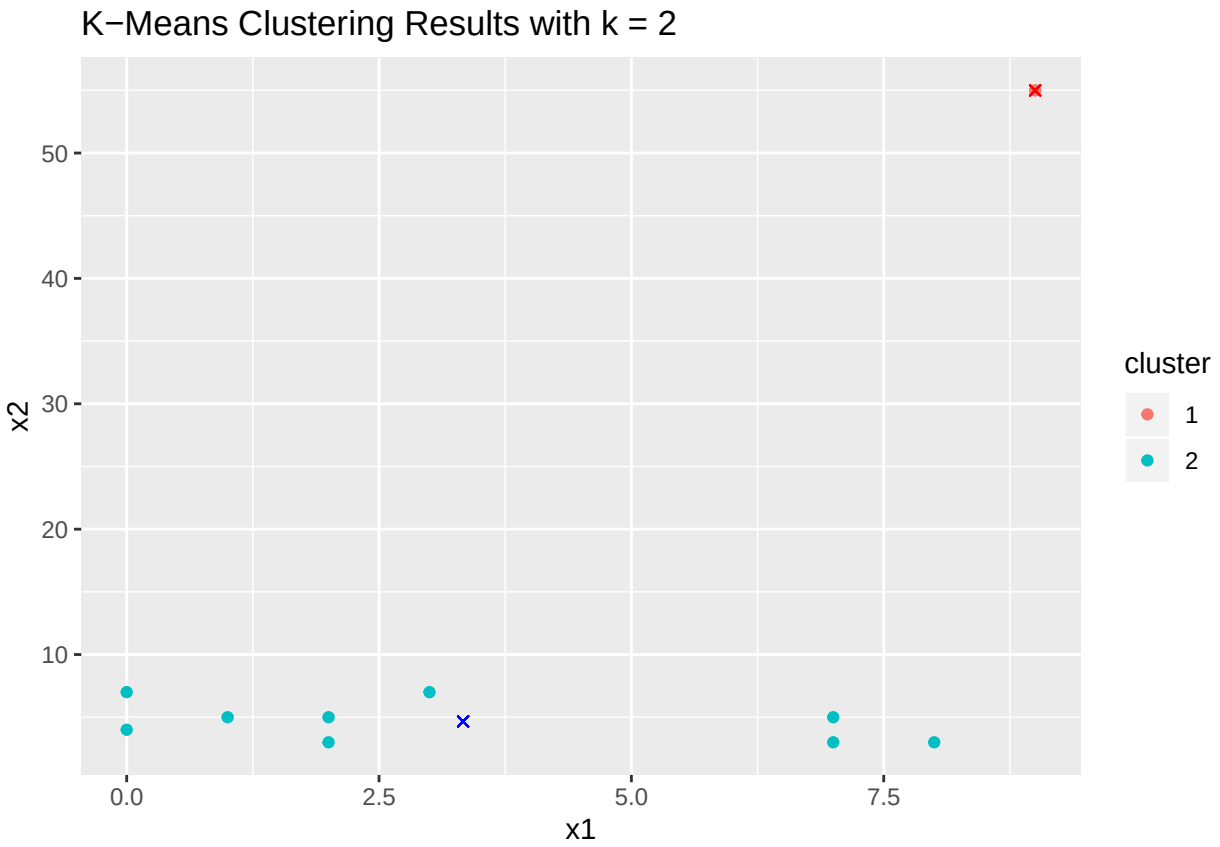
```r
for (i in 1:10){
  dis1 <- distance(dat[i,], centroid1)
  dis2 <- distance(dat[i,], centroid2)
  if (dis1 < dis2) {clust[i] <- 1}
  else{clust[i] <- 2}
}

(centroid1 <- c(mean(dat[clust == 1, 1]), mean(dat[clust == 1, 2])))
```

```
## [1]  9 55
```

```r
(centroid2 <- c(mean(dat[clust == 2, 1]), mean(dat[clust == 2, 2])))
```

```
## [1] 3.333333 4.666667
```

```r
dat %>%
  mutate(cluster = as.factor(clust)) %>%
  ggplot(aes(x1, x2, col = cluster)) +
  geom_point() +
  geom_point(x = centroid1[1],y = centroid1[2], pch = 4, colour = "red")+
  geom_point(x = centroid2[1],y = centroid2[2], pch = 4, colour = "blue")+
  ggtitle("K-Means Clustering Results with k = 2")
```



Problem 3

For this task you need to download World Value Survey (Wave6) data and try to understand the disposition of our country among others based on some criterias. The description of the variables and the survey are given with a separate file. Here is the link to obtain more information: . Choose the subset1 from Wave 6 data to perform the cluster analysis. Note that you need to use meaningful selections both of variables based on some topic/problem and countries.

(a) Describe thoroughly how and why you choose your subset of variables and observations. What is your goal? Hint: You need to prepare data for the next step.

(b) Use all(appropriate) tools/functions from our lecture to cluster the countries(both nested and untested algorithms). Interpret them.

(b1) Is your hierarchical clustering stable regards to between clusters distance measures?

(b2) Compare the results obtained from two different k-means.

(c) Make the conclusion (also based on cluster centers).

Solution 3

I have chosen the unswers about social values, attitudes,stereotypes and post soviet countries. Cause I want to see what countries have the same social values and what countries differ. But in the data there are only 10 post soviet countries. So I will work with 11 countries.

```r
dat <- readRDS("F00007762-WV6_Data_R_v20180912.rds")
dat <- dplyr::select(dat, c( "V2","V4", "V5", "V6", "V7", "V8", "V9", "V12", "V13", "V14", "V15", "V16"

subset <- dat %>% filter(V2 == 51|V2 ==31|V2 ==112|V2 ==498|V2 ==804|V2 ==643|V2 == 268|V2 == 440|V2 ==
unique(subset$V2)
```

```
##  [1]  51  31 112 233 268 398 417 643 804 860
```

```r
subset<- na.omit(subset)
set.seed(136)
dt <- subset %>% group_by(V2) %>% sample_n(1)

dt$V2<- factor(dt$V2, levels = c(31,51,112,233,268,398,417,643,804,860), labels = c("ARM", "AZR","BLR",
dt <- column_to_rownames(dt, var = "V2")
```

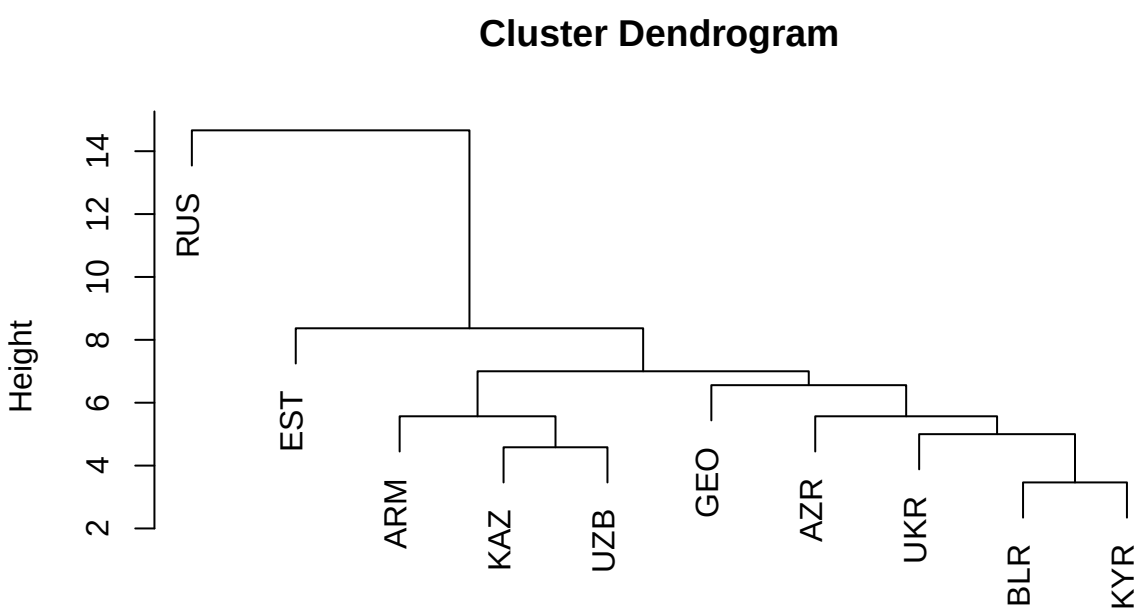
```r
km <- kmeans(x = dt,
             centers = 4)
km$cluster
```

```
## ARM AZR BLR EST GEO KAZ KYR RUS UKR UZB
##   2   4   4   2   3   3   3   1   4   3
```

```r
d <- dist(dt, method = "euclidian")
cl <- hclust(d,method = 'complete')

cl$height
```

```
## [1]  3.464102  4.582576  5.000000  5.567764  5.567764  6.557439  7.000000
## [8]  8.366600 14.662878
```

```r
plot(cl)
```




```
dt.clust <- cutree(cl, k = 4)
dt.clust
```

```
## ARM AZR BLR EST GEO KAZ KYR RUS UKR UZB
##   1   2   2   3   2   1   2   4   2   1
```